Report on

Visualization and exploration of Adult Dataset

CS245, Data Science Foundation, Chaojun Lu, Autumn 2017

叶泽林 515030910468

1 Introduction

Currently, data science is becoming ubiquitous in our society and showing its essentiality in many domains (e.g. medical industry [1, 2], finance [3], social media [4, 5]). With the rapid evolution and wide applications of data science, a series of efficient packages have been constantly developed [6, 7, 8, 9]. Utilizing these packages skillfully is of great importance nowadays. In this project, I tend to conduct an exploration over the *Adult* dataset and visualize the results with some of these packages.

2 Approaches

The *Adult* dataset is in the format of .CSV with 16281 lines, each line contains some basic information (age, job, gender and etc) of an adults. I explore and extract the information hidden in the data with the following steps:

1. Read the data with *pandas* [7] and reconstruct it as *DataFrame* type.
2. Select a target attribute (e.g. work time per week) to conduct analysis.
3. Select two attributes to explore the relation between them.
4. Visualize all analysis results with *matplotlib* [8] and *pandas*.

3 Experiments

3.1 Experiments Setup

Since the data is organized as .CSV format, it is easy to be recogbnized by *pandas*. I first construct the dataset as a *DataFrame* with 16281 lines and 15 columns, each line denotes the information of an adult while each column represents an attribute.

3.2 The Distribution of Single Attribute

Basically, it is of great importance to get the distribution of some attributes in the data. I hence explore some of them and visualize the results with different figure types.

1. The distribution of educational level in the dataset is shown in fig. 1(a).

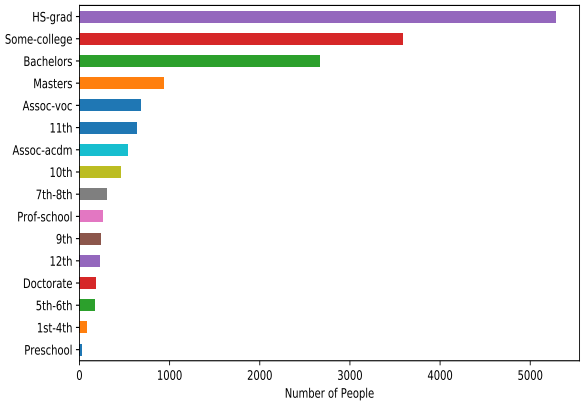
   The result shows the educational background of many adults are high school or college, occupying more than half of the proportion. Only a few people owns bad educational background(e.g. preschool).

2. The distribution of family relations can refer to fig. 1(b).

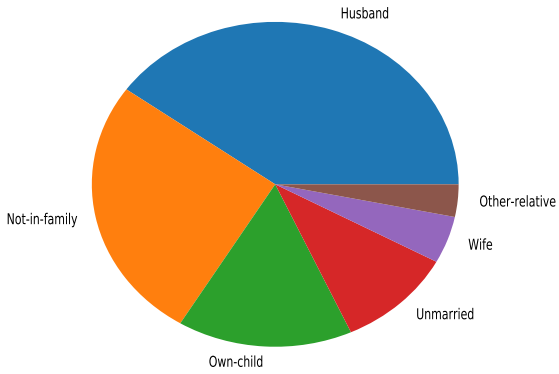
   The result indicates more than half of the adults have set up a family, while there still exist many adults remain unmarried or not in family.

3. The weekly working time distribution can be found in fig. 1(c).

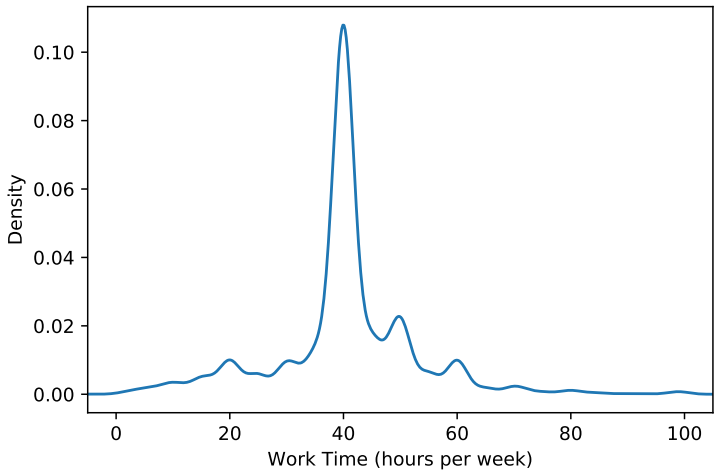
   This Kernel Density Estimation(KDE) figure shows that most adults work 40 hours per week, and a few people also work more than 60 hours per week. I will conduct some deep explorations in Sec. 3.3.1.



(a) Education Distribution



(b) Family Distribution



(c) Work Time Distribution

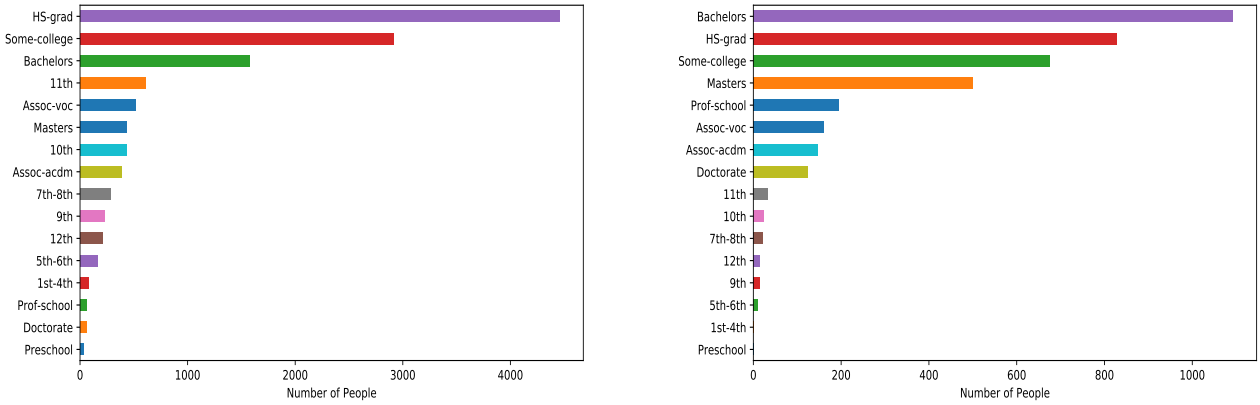
Figure 1: The Distribution of Some Single Attributes.

3.3 The Relation between Two Attributes

Generally speaking, it is not enough to extract only the distribution of single attribute. Therefore, in this section, I would explore some relations between two attributes.

3.3.1 Salary v.s. Educational Background

I tend to start with the relation between salary and educational backgound. The salary in *Adult* dataset is divided into two catagories: *more than 50K* and *less or equal to 50K*. The following two bar figures (Fig. 2) represent the distribution of educational background under different salary.



(a) Salary is less or equal to 50K

(b) Salary is more than 50K

Figure 2: The distribution of educational background under different salaries.

One of the main points implied in the two distribution is that higher education brings higher salary, which conforms to common sense.

3.3.2 Work Time v.s. Educational Background

I also compare the work time per week among three degrees: HS-grad, bachelor and master. (Fig. 3)

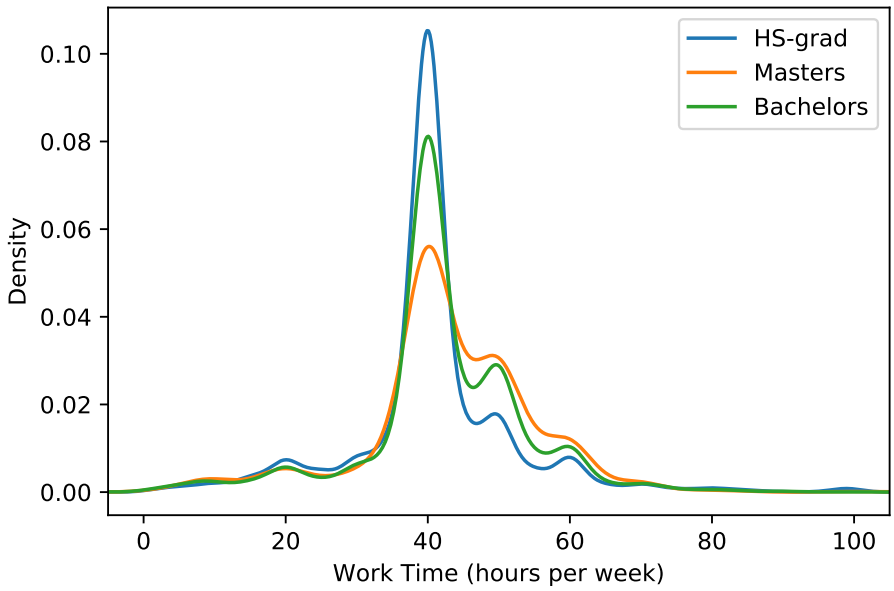


Figure 3: The work time per week among different degrees.

As is shown in Fig. 3, people with a higher degree tend to invest more time in their work.

3.3.3 Job Catagories v.s. Work Time

I am curious about the jobs that take people more than 50 hours a week. Therefore, I plot the distribution of such jobs (Fig. 4).

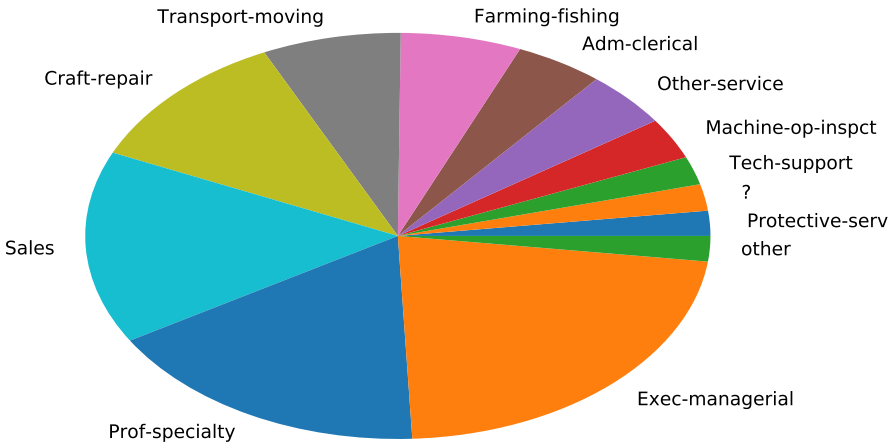


Figure 4: The job catagories that take people more than 50 hours a week.

Different from my intuition, the result indicates that time-consuming jobs are focused on management and research fields, instead of manufacturing industry.

4 Conclusion and Discussion

In this project, I carry out the visualization and exploration of *Adult* dataset and discover some interesting phenomenons reflected from it.

According to the results of above analysis, the distribution of each attribute has some kinds of relation with others. Thus, I think I can train a model to decode this relation as some weights and predict target attributes from others in my future learning process.

Ultimately, I tend to express my sincere thanks to Professor Chaojun Lu for his patient explanation and guidance in lectures! Thank you!

References

[1] S. P. Bhavnani, D. Muñoz, and A. Bagai, "Data science in healthcare: implications for early career investigators," *Circulation: Cardiovascular Quality and Outcomes*, vol. 9, no. 6, pp. 683–687, 2016.

[2] Y. Liang and A. Kelemen, "Big data science and its applications in health and medical research: Challenges and opportunities," *Austin Journal of Biometrics & Biostatistics*, vol. 7, no. 3, 2016.

[3] S. O' Halloran, S. Maskey, G. McAllister, D. K. Park, and K. Chen, "Data science and political economy: application to financial regulatory structure," *RSF*, 2016.

[4] W. Xiao-fan, "Data science and social network: Big data, small world," *Science and Society*, vol. 1, p. 003, 2014.

[5] C. Vande Kerckhove, *Data science for modeling opinion dynamics on social media*. PhD thesis, UCL-Université Catholique de Louvain, 2017.

[6] "Numpy." http://www.numpy.org/.

[7] "Pandas." http://pandas.pydata.org/.

[8] “matplotlib.” https://matplotlib.org/.

[9] “scikit-learn.” http://scikit-learn.org/stable/.